DOCUMENT RESUME

ED 234 056 TM 830 274

| | |
|---|---|
| AUTHOR | Boldt, Robert F. |
| TITLE | Status of Research on Item Content and Differential Performance on Tests Used in Higher Education. |
| INSTITUTION | Educational Testing Service, Princeton, N.J. |
| REPORT NO | ETS-RR-83-3 |
| PUB DATE | Feb 83 |
| NOTE | 42p. |
| AVAILABLE FROM | Educational Testing Service, Research Publications, R116, Princeton, NJ 08541 |
| PUB TYPE | Reports - Research/Technical (143) -- Information Analyses (070) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | Classification; Cultural Differences; Group Behavior; Higher Education; *Minority Groups; *Predictive Measurement; Racial Differences; *Research Methodology; Sex Differences; *Test Bias; *Test Items |
| IDENTIFIERS | Discrepancy Analysis; *Outliers; *Test Content |

ABSTRACT

"Outlier studies" identify items for which extreme differences in performance by contrasting groups occur; these extreme items are the "outliers" referred to. Review of the studies conducted on tests receiving major use in higher education reveals that though one cannot make a priori classifications of outliers with confidence, one can with reasonable confidence predict the relatively advantaged group for many verbal items if they subsequently prove to be outliers as follows: aesthetic-philosophical, human relations or female oriented content relatively favors females as opposed to males; practical affairs, science or male oriented content relatively favors males as opposed to females; science content relatively favors whites as opposed to blacks. For test content that varies in the degree of relatedness to minorities, one would predict a relative advantage for those outliers that are most related to minorities. The magnitude of the differences found is not large; perhaps larger differences would be found if classifications other than race and sex, which are the most common, were used. It has been found that differences in cultural or national origin produce larger discrepancies in item difficulty than differences in race or sex of essentially native American groups. (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

RR-83-3

ED234056

RESEARCH REPORT

# STATUS OF RESEARCH ON ITEM CONTENT AND DIFFERENTIAL PERFORMANCE ON TESTS USED IN HIGHER EDUCATION

Robert F. Boldt

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. C. Weitzenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

February 1983

Educational Testing Service
Princeton, New Jersey

TM 830 274

Copyright © 1983, Educational Testing Service. All rights reserved.

ABSTRACT

"Outlier studies" identify items for which extreme differences in performance by contrasting groups occur; these extreme items are the "outliers" referred to. Review of the studies conducted on tests receiving major use in higher education reveals that though one cannot make a priori classifications of outliers with confidence, one can with reasonable confidence predict the relatively advantaged group for many verbal items if they subsequently prove to be outliers as follows: aesthetic-philosophical, human relations or female oriented content relatively favors females as opposed to males; practical affairs, science or male oriented content relatively favors males as opposed to females; science content relatively favors whites as opposed to blacks. For test content that varies in the degree of relatedness to minorities, one would predict a relative advantage for those outliers that are most related to minorities. The magnitude of the differences found is not large; perhaps larger differences would be found if classifications other than race and sex, which are the most common, were used. It has been found that differences in cultural or national origin produce larger discrepancies in item difficulty than differences in race or sex of essentially native American groups.

## INTRODUCTION

The test item performances of groups that differ by race, sex or other socially relevant characteristics have long been of scientific and practical interest. In recent years such work has frequently focused on "outliers" in analyzing the performance of such groups on tests. By "outliers" is meant items for which group differences in performance markedly exceed those of most items in the test or sub-test in which the outlying items are imbedded. The differences on which the identification of outliers has been based have been called "performance differentials" and "performance discrepancies." Carlton and Marco (1981) have identified 19 analyses of differences in item performance; Linn & Harnish (1979) and Linn, Levine, Hastings, and Wardrop (1981) offer examples of studies of differential item performance.

Three motives for such studies can be identified: one motive is to arrive at some explanation of observed group differences in performance; a second is to identify sources of test score variation that are irrelevant to the quality for which measurement is needed so they can be eliminated; a third is to minimize the unfairness of tests to particular groups of people. Each of these purposes might be served by identifying categories of item content that are associated with the particularly good or bad performance of any particular group relative to another. The identification of item content associated with group performance discrepancies on particular items might lead to fruitful speculation as to why that content favors a particular group, perhaps suggesting what, or whether anything, should be done about it. In particular, if the source of

performance variation by groups is irrelevant to the purpose of measurement, it can be eliminated. Indeed, to the extent that one can modify the content of a test without affecting the relevant measurement aspects or the intellectual task it poses, one can choose content that balances out the unfair advantage of particular groups (Donlon, 1973; Medley & Quirk, 1972; Flaugher & Schrader, 1978).

The item categories of interest here are not those by which items are currently classified into tests or sub-tests. Group performance differences in the usual test or sub-test types are well known. Knowing about them has not led to their control, and it is scientifically and socially unacceptable to leave it that the performance differences exist "because of" sex differences or racial differences. It is certainly scientifically unwarranted to associate the differences observed with hereditary effects or to assume that they are environmentally immutable simply because we have discovered no plausible explanation for them. We therefore turn to the examination of within test or sub-test differences in item performance to formulate hypotheses about group differences in test performance.

The purpose of the present project was to review reports on outliers and related reports in order to summarize their implications relating group differences and item content. This review is confined to studies of tests that receive major use in higher education. The discussion uses aptitude and achievement as the major category heads. Despite the fuzziness of this distinction in many contexts, aptitude and achievement

can be clearly enough separated in the studies of interest here. Before beginning the aptitude section of this report, however, it will be useful to give a very brief review of the research methods involved. Shepard, Camilli, and Averill (1981) give more extensive discussion, bibliography and a study of these methods, which they call "procedures for detecting test item bias."

## METHODOLOGICAL BACKGROUND

Several definitions of outliers were used in the studies reviewed here; they do not necessarily identify the same items (Stricker, 1981). These definitions may be differentiated by the type of item statistic examined to establish the extent of the performance discrepancy, and by the type of inferential procedure used, if any, to separate outliers from the rest of the items.

The item difficulty, P , equal to the fraction of people passing the item or a transformation of that fraction, is the most common statistic examined to express differential performance. The most commonly used transformation of P is Delta, which is the probit of P using a central tendency of 13 and a standard deviation of four. The Arcsin transformation has also been used. A plot of P against Delta produces the familiar ogive; a plot of P against Arcsin P bears a very high resemblance to the ogive. A virtue of these transformations is that their use tends to linearize bivariate plots when the axes of the plots are defined by the transformed difficulties. Indeed the linear relationship is so good that Echternacht (1972) proposed setting up confidence bands based on an assumed joint normal distribution of the Deltas; items falling outside the bands would be considered outliers.

Item difficulties are probably used so often because of their close intuitive relation to the average group performance. In fact, except for conventional allowances for drop-outs, omits and guessing, the average test performance is within a linear transformation of the sum of the item difficulties. Thus, average test performance for a group tends to be monotonically related to the sum of Ps or their transformations.

The item-test score correlation coefficients are sometimes studied also. Differential performance on these correlations indicates a discrepancy in the discriminating power of an item in one group, as opposed to another. In a rough sense, comparisons of such correlations index the extent to which the item (or test) measures the same thing in one group as it does in another.

Measures of difficulty or of item-test association have been the elements of many of the comparisons that have led to identification of items as outliers. In a simplest procedure, one calculates for each group the statistics on which comparison is to be based and takes their differences. The items for which the differences are extreme are the outliers. This procedure could also be accurately described as follows: Generate a bivariate plot, where items are the points and the axes are the values of the item statistics computed using the groups being compared; put a $45^{o}$ reference line through the origin of the plot; identify as outliers those items located farthest from the reference line. A more common procedure is to use, for each group, the deviation of item statistics from the group mean, in effect passing the $45^{o}$ line through the centroid for the plot. This procedure was used by Coffman (1961), Donlon (1973) and Levin (1970). The procedure of the plot reflects the variation in performance as referenced by performance on the rest of the items. Angoff and Ford (1973) further modified the procedure, suggesting the principle axis of the plot as the reference line, rather than the $45^{o}$ line through the origin or centroid, and gave equations for the line and distances from the line. Most studies have used Angoff and Ford's approach with item Delta values of the different groups

defining the reference axes. Using the distance from the principle axis of the item plots will be referred to hereafter as the Delta-plot method.

In using the major axis or the centroid as the reference line, overall test performance modifies the definition of outliers. Indeed, an item that plots almost directly on a $45^{o}$ line through the origin would be identified as an outlier if the performance of the two groups were sufficiently different in the rest of the items (as happened to Donlon in his 1973 study of mathematics items). Some other methods of identifying outliers also take total test performance into account in some way. One of them is to produce matched samples by selecting a sample from the largest group that has the same total score distribution as that of the smaller group. Then the $45^{o}$ line can be used as the reference line (Cowell, 1968). Item Delta equating (Hecht & Swineford, 1981), which computes the Deltas on different groups but then adjusts them for population differences and uses the $45^{o}$ reference line, was applied by Conrad & Wallmark (1975). Because in these methods the advantage of a group on an item is defined relative to the common trend rather than relative to the raw difference, the term "relative advantage" is used repeatedly in the paragraphs to follow to emphasize that the differential item performance is evaluated in the context of the performance on all items.

In a correlational approach, Stricker (1981) explicitly uses the total test score to control the interpretation of performance differences. Stricker gets a partial correlation of item performance with race controlling total test scores so that, in a linear system, he indexes the association of race with item performance for people of the same ability

level. Other methods allow relaxation of the linearity requirement by controlling on total score using a group by score level contingency table (Alderman and Holland, 1981). In this approach, which was proposed by Scheuneman (1979), one hopes that by keeping score intervals narrow enough, group differences in total score within the intervals will be minimal, and in the absence of a group effect the proportion pass would be about the same. Various indices of the success of this hypothesis have been proposed.

Rather than use the total score, Wightman (1979) and Marco (Lord, 1977) have used the ability inferred in an item response theoretical approach. They constructed separate item response curves for the groups, and then conducted significance tests to see if the parameters were the same within statistical variation. Content interpretations by these authors are not available; Lord (1977) comments that interpretation of Marco's results has not been successful. Stricker (1981) also used item response theory and obtained some relationship with context. He did not, however, use the item response curve for interpretation of results. Such results are not, therefore, available for this paper.

The paragraphs above attempt to indicate the various methods of identifying outliers in a way that maintains some rough conceptual relationship among them. The methods are not essentially interchangeable, though they are undeniably related. Nor is any great preference for a particular method implied. Rather, the methods have their particular advantages and disadvantages that have yet to be sorted out. That sorting excedes the scope of this paper.

One common feature of the methods is the use of some numerical standard by which the degree of departure from a common standard is measured. For some methods the standard is the distance from a reference line, for others it may be a chi-square value or a significance level. It is computed for all items in the test or sub-test under study and can be used with a cut score that separates outliers from the rest. For most of the studies referenced in this report, the choice of cut scores is based on one of two types of rationale: statistical significance or the number of items identified as outliers. Where the sample sizes are large, many outliers are found because of the great power of significance tests conducted with many replications. Indeed Donlon (1973) and Stricker (1981) all used significance tests with large samples and found many "outliers." We would expect Marco (Lord, 1977) and Wightman (1979) also to find many outliers. In such cases the cut score might be very close to the average value of the numerical standard for all the items on the test. This result merely stresses the fact that evaluation of the social significance of group discrepancies in performance requires more information than the significance test provides. Angoff (1981) has concluded that bases other than traditional significance tests should be sought to evaluate differential performance in outlier studies.

## APTITUDE

Many partitionings of candidate populations are possible, of course, but the usual context of outlier studies is that of bias in testing. Hence we have results only for sex and race groupings. When the two principle types of academic aptitude variables, verbal and mathematics,

are cross classified with the two groups four combinations result. These combinations give the groups of aptitude studies that are covered.

Verbal-Sex

Male/female differences in test performance have long been recorded in the differential psychology literature. Traditionally women are regarded as more verbal, men as more quantitative. However, the differences are not uniform across items. Coffman (1961), based on an examination of extreme PSAT item differences, hypothesized that items related to "people" might favor women relatively more; items related to "things" might tend to favor men. He speculated that the test could be manipulated to produce relatively better performance for either sex. Donlon (1973) noted that the advantage of females in verbal performance seems to have disappeared. He conducted an analysis like that of Coffman using data for a 1964 SAT administration. Like Coffman, he examined the items with the most extreme differential performance and felt that the Coffman surmise was supported. Straussberg-Rosenberg and Donlon (1975), using the Delta-plot method, obtained results that supported those of Coffman (1961) and Donlon (1973) in that the deviant items that favored males tended to be "thing"-related, and those that favored females tended to be "person"-related. Straussberg-Rosenberg and Donlon further elaborated the "people"-"thing" principle by identifying the test developer categories of "world of practical affairs" and "science" with "things," and "aesthetic-philosophical" and "human relationships" with "people." The definitions of these categories are given below:

A. Aesthetics-Philosophy items deal with art, architecture, literature, drama, music, religion, philosophy;

B. World of Practical Affairs items deal with economics, government, history, politics, transportation, communication, sports;

C. Science items deal with research, mathematics, agriculture, engineering, medicine, weather, manual arts, inventions, geography, psychology;

D. Human relationships items deal with interpersonal relationships character analysis, emotions, family.

Items classified according to these test development categories tend to deviate from the major axis of the Delta plot in the expected direction. Reference to male-female characters seemed not to favor either sex.

Using SAT items in another analysis like that of Straussberg-Rosenberg and Donlon, Stern (1978) supported the finding that aesthetic-philosophical and human relationship items favored white females, relatively, while science and practical affairs items relatively favored both black and white males.

Donlon et al. (1980) studied sex differences on the Graduate Record Examinations Aptitude Test (GRE-AT). This study also used the Delta-plot method. The results were generally supportive of the previously mentioned hypothesis for science, practical affairs, and human relationships, though one item, dealing with practical affairs, ran contrary to the hypothesis. Conrad and Walmark (1975) also found a slight tendency for items on science reading passages on GRE-AT to give relative favor to males.

Contrasting with the above results are those of Cowell and Swineford (1972). Using the Delta comparison method to study sex differences on Law School Admissions Test (LSAT) items, they found essentially no

instance of interpretable item bias. Rather than using some kind of significance test, they came to this conclusion by examining scatter plots of item difficulties and noticing that the plots were very tightly distributed around the major axes. They pointed to an aspect of their data that applied to those of other testing programs as well--that cross-sex correlations of item difficulties tend to be in the very high nineties. Thus, departures of Delta from the major axis are actually very small in numerical magnitude. Francesco (1975), in an analysis of variance, found that from 80 to over 90 percent of variation in transformed item difficulties was due to items alone; item by sex interactions, though significant, counted for only one to three percent. Francesco implies, however, that a more sensitive examination of differences would have led Cowell and Swineford to a different conclusion. In particular, Francesco felt that the use of significance tests rather than examination of the plots would have identified outliers.

That sex differences in LSAT item performance are related to item characteristics was demonstrated in the unpublished study by Francesco (1975), who correlated sex differences in difficulty with a number of rated characteristics of items. Few of her correlations were significant, but the traditional finding that math favored males and verbal favored females was observed, and the inclusion of content on business, work, and money favored males.

Stricker (1981) also examined GRE-AT and, with a somewhat different significance test, found very few items that differentially favored males or females in terms of item difficulty. He supplemented the test development classifications with two categories for explicit reference

to blacks and females, but found no significant difference. He did, however, find significant content differences on the partial correlation of items with sex, controlling on total score. In these correlations the aesthetic-philosophical, human relationship, and female reference partial correlations tended to be positive, while those on the world of practical affairs and science tended to be negative. Positive partial correlations indicated a favoring of females, hence these results seem to support the previous ones, although counter examples occurred with some frequency in Stricker's data.

Sinnott (1980) analyzed sex differences in item performance on Graduate Management Aptitude Tests (GMAT). In her study the females out-performed the males on verbal materials. Her finding on national origin is mentioned in the achievement section of this report under national origin.

Verbal-Race

As with male-female differences, race differences in verbal performance on tests have long been known. Using analysis of variance at the item level, Cardell and Coffman (1974) noted item by race interactions, as did Cleary and Hilton (1968). Angoff and Ford (1973) used the Delta method, which was also used by Cowell (1968) in an early unpublished study of the February 1963 Admission Test for Graduate School of Business (now GMAT). Cowell found a number of items on which performance was relatively discrepant, but no interpretation of these differences was offered.

A series of analyses of item performance discrepancies by race has been conducted using the Delta plot method with Scholastic Aptitude Test

items. As regarded one of these analyses, Stern's (1974) only finding on content was that two items that favored whites were based on science reading passages. No verbal content effect was reported in a later analysis (Stern, 1975), but Cook and Stern (1975) report that narrative and "minority relevant" reading passages were associated with reading comprehension items that were relatively easier for black candidates. Stern, in 1978, found a minority relevant passage easier for black males, also finding that an argumentative reading passage favored blacks and a humanities passage was relatively difficult for black females. Blew and Ishizuka (1978) and Blew and Stern (1979) reported additional analyses of SAT items but no content differences were noted. Examination of all these studies reveals that (1) when a science reading comprehension outlier is found it is without exception relatively more difficult for blacks, (2) when a social studies reading outlier is found it is without exception relatively easier for blacks, (3) average deviations from the Delta plots are consistent with the outlier findings (1) and (2), and (4) "minority relevant" items are found only in humanities and social studies items.

The finding that items based on passages judged minority-relevant were relatiely easier for blacks was repeated by Conrad and Wallmark (1975) using the GRE-AT.

Stricker (1981), using his partial correlation of race with item performance controlled on total score, found that for males, aesthetic-philosophical and human relationship content favored blacks relatively; world of practical affairs and science favored whites relatively. In

contrast with Cook and Stern's (1975) results, Stricker's data do not yield significant partial correlations associated with the presence of "black content."

Math-Sex

As with SAT-verbal items, Donlon (1973) studied the difficulty differences of SAT math items for males and females. He found that, in the test examined, the items that seemed purely algebraic, as opposed to those with some story or real world content, gave males less of an advantage than did those with verbal content. Donlon, Ekstrom and Lockheed (1979) have found that verbal content tends to be masculine oriented, and that masculine orientation is related to relatively better performance by males. Therefore the relative advantage of females on the purely algebraic material might not be due to the nature of the mathematical processes involved but to the sex orientation of the language. Indeed, to give an exception that supports the rule, items about shopping and the laundry, which are thought to be topics more closely related to women, relatively favored females.

Straussberg-Rosenberg and Donlon (1975) analyzed the SAT using the Delta plot method, and found that geometry and arithmetic items were relatively easier for males; algebra, elementary number theory and letter addition (filling in missing digits in multiple addendum addition problems) were relatively easier for females. They also found that real world reference items tended relatively to favor males.

In contrast to the SAT results by Donlon and others, those by Stern (1978) do not reveal a consistent outlier difference in item performance by sex, nor do those of Conrad and Wallmark (1975), and

Stricker (1981). Indeed in the report by Donlon et al. (1980) of GRE-AT math items, the content interpretation of outliers is not evident.

Sinnott (1980) found some significantly discrepant performances in the GMAT. She reported a tendency for word problems in problem solving items relatively to favor men. One nonword problem that favored men was a geometry item.

Math-Race

As with verbal content, race differences have long been a known aspect of mathematical test item performance, but few content oriented factors that contribute to these differences has been reported. In an early study of the ATGSB, Cowell found a tendency for items involving percentages to be relatively more difficult for blacks, a finding that was not repeated in Sinnott's (1980) study of the GMAT. Neither was it repeated in an analysis conducted for CEEB in which Braswell (Cook & Stern, 1975) examined items identified as producing discrepant performance levels between blacks and whites. However, Braswell did note that four of six outliers dealt with ratios. In a subsequent CEEB analysis (Stern, 1978) four out of five items that were relatively harder for black females than white females dealt with decimals or fractions, though Stern did not list this trend as a finding of the study. No other content inferences were given in this series.

In an analysis of data for a GRE-AT administration, Conrad and Wallmark (1975) found that blacks have relatively more difficulty with items that required the structuring of solutions to problems by translating words to algebraic expressions; items with relatively less difficulty

for blacks required more straightforward application of quantitative processing.

## ACHIEVEMENT

Though many types of academic achievement tests exist, the number that have been the subject of outlier study is few. Results are available in Verbal and English achievement for various language background groups. Though there is a large literature on sex differences in testing, outlier studies on sex differences in achievement are available only in the physics area (Lockheed, 1982). Finally, there are several studies where the achievement test items performance of racial groups is compared.

### Language of National Origin

Differential item performance that is associated with the language of national origin is of particular interest in the Test of English as a Foreign Language (TOEFL), in part because the studies contribute evidence of construct validity of the test, which is used in foreign student admissions. Correlations between item difficulties suggest that differences across languages exceed those found across sex or race in the same language. It will be remembered that across the groups examined for verbal and math aptitude in English, the vocabulary item Delta correlations were in the 90's. In contrast, Angoff and Sharon (1974) found a correlation of .73 for Delta of Spanish examinees, one of .88 Deltas of Gujarati examinees, against Deltas of a sample from the general TOEFL population, which is a mixture of these and other language groups. Comparable correlations for German, Arabic, Chinese and Japanese test-takers were

intermediate to the Spanish Gujarati values. Also, Alderman and Holland (1981) studied TOEFL item difficulty intercorrelations for six language groups: Germanic, Spanish, African, Chinese, Japanese and Arabic. These correlations were reported by section for two administrations and range from .43 to .93 with a median value of .80. Alderman and Holland (1981) also drew two Chinese samples for each administration; the correlations of Deltas for both pairs of samples did not drop below .99 for any item type. Clearly, discrepant performances are induced by differences in language of national origin.

Alderman and Holland (1981) attempted to discover a linguistic explanation of the discrepant performances by asking specialists in English as a second language to examine the items, distractors and item statistics from one administration, then to formulate some linguistic explanation of the differences, and finally to apply the principles in an attempt to predict those items that will produce discrepant performances in a second administration. Unfortunately, there are no principles to report because the attempt was quite unsuccessful.

Sinnott (1980) studied discrepant item performances for foreign GMAT examinees who claimed fluency in Chinese, English, French, Indo-Iranian, Japanese or Spanish. She also obtained performances by a sample of examinees that claimed U. S. citizenship. Sinnott found that one passage, which was answerable if one were familiar with the times of Roosevelt's New Deal, was relatively easier for Japanese examinees. She speculated that the reason that the other groups did not find the New Deal items especially easy was that they were not as well acquainted with that period of American History, or that their command of English as a

group was good enough that they needed no supplementary knowledge to read the paragraph effectively. No trends were uncovered in the section on practical business judgment. In the English usage section Sinnott found items that were relatively easy for foreign examinees and that tested basic principles of language--principles whose usage is prone to improvement by drill. Indeed, the foreign examinees did as well or better than the U. S. candidates on some of these items. Sinnott's discussion suggests that foreign candidates deal with the English language in a formal way, and are not affected by awkward but correct phrasing.

The use of GMAT allowed Sinnott to present the only comparison of foreign population on math item difficulties. Her principle finding was that if the item dealt with real world concepts expressed through language the foreign candidates had more difficulty than with symbolically expressed problems.

Achievement-Sex

It has already been pointed out that when items are found on which discrepant performances by the sexes occur, science content is often involved. This finding could be sharpened by a study of outliers on physics tests. Such a study was conducted by Wheeler and Harris (1981), in which data from a CEEB physics achievement test were analyzed. The test covers mechanics, electricity and magnetism, optics and waves, heat and kinetic theory and modern physics. Males out-performed females on the test, though the authors show that the performance is related to the amount of preparation in physics, which favors males. Unfortunately, adjustments for math aptitude were not available to the authors because

of limited funds. Delta plot analyses failed to detect any striking tendencies for particular physics content areas to produce outliers; except for modern physics and electricity and magnetism, which favored males, the content categories produced no outliers or produced them for both sexes. Modern physics produced only one outlier, and electricity and magnetism produced three. The Wheeler and Harris (1981) study is the only achievement test study of sex differences that was found.

Achievement-Race

The items of the Common Examination part of the National Teacher's Examination (NTE) have been analyzed in a study by Levin (1970), who noted that NTE Commons examination developers had made four assumptions about test performance on literature items by blacks and whites:

"1. Questions on black literature, questions dealing with black artists and musicians, and questions dealing with the black experience will be easier for black students.

2. Questions calling for the analysis of given stimulus materials and questions calling for an understanding of material actually presented to the student, will be easier for black students since their preparation can be assumed to be less thorough in factual material than that of his white counterpart.

3. Questions that rely heavily on factual recall and that deal with earlier literary materials (Chaucer, Emerson, Donne, Greek myths, for example) will be more difficult for black students, again, because we assume their preparation to be less thorough in conventional and traditional literacy materials.

4. Questions dealing with contemporary materials, 'with-it,' "relevant" literature, will be relatively easier for black students who might be assumed to be particularly aware of American social problems and interested in material treating such live issues."

Levin noted,

"Based on these beliefs a group of items about black authors or black experiences were included in the test. As measured by the average delta, these items proved easier for the 145 black students from southern Negro colleges than for 200 white students from southern colleges. Thus this small study bears out the first assumption above. It was the only one supported by the analysis, as will be seen. [Levin found that] "items on Shakespeare, Chaucer, Oedipus, Milton, Arnold, Donne, Keats, Wordsworth, even the Rivals, and questions in chronology and genre proved relatively less difficult for black students. In American literature these students compared favorably on questions dealing with such figures as Emerson and Thoreau. Thus statistics on these items suggest that black students receive a more conventional and traditional literary education. Literary 'Giants,' and major documents and movements would seem to be emphasized and literary history must have a significant place in their training. Items which prove relatively more difficult dealt with Dylan Thomas, Browning, Byron, Lamb, Shaw, Tennyson, Williams, Wilder, even T. S. Eliot. The students proved weak in literary analysis. It may well be that

colleges emphasizing literary history tend to scant close reading, new criticism, and intensive examination of the literary text. Certainly, the latter represents a more modern approach to literature, and colleges that train these black students may simply not have caught-up in this sense.

It is also quite possible that test questions demanding verbal facility, skill with dealing with nuance and drawing inferences are more difficult because verbal manipulation is the area where disadvantaged students appear to be weakest. Whatever the explanation, the item statistic suggests that students from black colleges are able to perform better when asked to deal with factual material dependent upon recall if that factual material centers upon the most traditional elements of literary study. [Levin noted further that the black students tended to find] historical questions dealing with structural descriptions especially difficult, perhaps because the focus of these questions is not in accord with the more traditional kinds of preparation they have had. Instruction in the history of the language has become much more pervasive as required course for teachers in the last 20 years The field is understaffed, and stiff competition for available linguists is still a problem in academia. If black colleges cannot meet this economic competition, they may not be providing modern language-linguistic instruction. The generalization regarding historical questions dealing with structural descriptions seems to be supported by black students' scores on language structure. In

this section items are couched in the traditional terminology of grammar and the figures suggest that this material is more familiar to black students. Composition and rhetoric, and to a lesser extent the structure items require the student to deal with language in context and to abstract from a verbal situation an appropriate description or evaluation.

The test includes the small but significant number of items on contemporary literature. It is interesting to note that there is no evidence that black students are particularly strong on such materials as Catcher in the Rye, or Portnoy's Complaint. If they are 'with it' it is in relation to 'with it' material that speaks to their condition, not the literature of middle-class hang-ups. Tolkien, Vonnegut, Tennessee Williams, film directors and Orwell are relatively more difficult for them."

With this result, three of the four hypotheses are contradicted and the definition of "black content" was formulated as content that, while difficult for most examinees, is specifically related to black examinees as well as being credibly related to the academic field with which the examination deals.

Humphries (1979) studied the relative performance of blacks versus the total group of NTE Commons Examinations takers. This examination has several sections. Items within each of several categories in each section were classified on judgmental grounds as having highly similar content. Humphries calculated the mean percents pass for the items in each content category. When these percents are transformed into Deltas and examined one notes that science was relatively difficult for

the blacks and social studies relatively easy. Some difficulty in interpreting this study is typified by a result in the Professional Education area. One of the contents in Professional Education deals with history, philosophy and social development relative to education. This content was relatively most difficult for blacks as compared with the other contents under Professional Education. However, when the Professional Education items were seperately examined via the Delta plot method, the outliers were all from other contents! Hence the interpretation of the results of this study is quite unclear. Even so, the attempt to formulate and evaluate content categories that may be related to differential item performance is a highly desireable feature of this study.

## DISCUSSION

The existence of a number of outlier studies of tests commonly used in higher education invites an integration of results; the patterns of confirmation and contrast may lead to broader conclusions or conjectures than could be reached by focusing on any one test or test program. Granted that different methods and authors are involved, still it is likely that any existing strong basic variance in differential performance produced by content variation should come through. One looks for patterns of results that might suggest experimental or survey research to conduct critical tests of hypotheses based on the conjectures, including research on the social processes that might be responsible for item performance differential. One attempts to turn hindsight into foresight for subsequent forms to emerge from the test development process. Conceivably curriculum recommendations could result.

The implications to be drawn here must, however, be constrained by the fact that no content classification has been discovered by which one can identify an outlier with confidence. Estimates of the actual magnitude of the effects for any particular classification have been infrequently made, in part because each classification includes some non-outliers, which are not examined in outlier studies. With more extensive studies of the content classifications that have been used to date one would expect content effects to be significant but modest in size. This expectation exists because the correlations between Deltas for groups for whom English is essentially the basic language tend to be in the nineties. These large correlations leave little room for deviation from the major axis of the Delta plots. But where language differences are involved the opportunity for differential performance is much greater. It has been pointed out that lower correlations obtained when Deltas for items from various language groups taking the Test of English as a Foreign Language are compared. The lowest correlation between Deltas, .60, was observed by Angoff and Modu (1973) who administered Spanish- and English-language versions of the same items to Puerto Rican and American students respectively, and then correlated the resulting item Deltas. Unfortunately, no reliable content effects on the differential item performance of groups with different language backgrounds have been catalogued as yet.

Though one cannot make <u>a priori</u> classifications of outliers with confidence, one can with reasonable confidence predict the relatively advantaged group for many items if they subsequently prove to be outliers. For verbal items with aesthetic-philosophical, human relations or female

oriented content one would predict a relative advantage for females as opposed to males. For verbal items involving practical affairs, science or male oriented content one would predict a relative advantage for males as opposed to females. For verbal items with science content one would predict a relative advantage for whites as opposed to blacks. For test content that varies in the degree or relatedness to minorities, one would predict a relative advantage for those outliers that are most related to minorities.

These findings do not support the notion that "bias" is what outlier studies discover. True, certain of the "female-" and "minority-related" items introduce content that seems clearly extraneous to the purpose of the test. But other content categories can be regarded differently. For example, the existence of a relative disadvantage for blacks on physical science oriented material doesn't necessarily mean that verbal items should not be administered with physical science content; too little is known about aptitudes to make that decision, though excision of the material is one alternative to consider. What the finding does point up is an educational deficit that might be worth addressing further. Clearly the findings of Wheeler and Harris (1981) do not establish that certain facts regarding modern physics and magnetism should be struck out of a physics achievement examination because females exhibited a relative disadvantage when asked about them. Rather the subject is what it is, but there are some educational problems as regards women.

One must also keep in mind that the content categories refer to content encountered in test items and should not be generalized beyond

that context. Consider, for example, the finding that human relations relatively favors women and that practical affairs such as business relatively favors men. These findings might seem strange to students of management and business, who learn that interpersonal relations will be extremely important in their futures, and who might conclude that skill in human relations is essential to much that is practical in affairs. To them, success in practical affairs entails success in human relations contrary to what the test items might seem to imply. While there is probably a way to rationalize this apparent contradiction, one might nevertheless learn from it that the translation of regularities "in the world" to principles concerning the effects of test item content on differential item performance, as well as in the other direction, is not straightforward. Theories about what makes tests "biased" are probably oversimple.

In the paragraphs above it has been noted that the outlier studies, which originated in the context of bias research, don't necessarily study bias and may not be able to discover large effects even if those effects prove reliable. To avoid being overcritical, note that the outlier studies were reasonable to do and would have yielded a considerable payoff for a very modest investment if the results had been more substantial and interpretable. To make progress, though, it seems that all of the traditional scientific stages are needed: collection of anecdotes, formulation of hypotheses, survey and experiment, and theory construction and confirmation. We seem currently to be in the anecdotal stage having some difficulty in the formulation of hypotheses. We are in the anecdotal

stage in that we are observing the pertinent events under uncontrolled conditions afforded by the use of operational tests which are not designed to test these hypotheses. Outlier items, each with many characteristics, are identified with the one characteristic that happens to be of interest in a particular study being selected as the reason for the variation, i.e., the "cause" is selected and the anecdote is completed. But for the most part, the enunciation of generalized principles and the evaluation and modification of those principles through survey and experimentation has not occurred. Indeed, mechanisms have been applied to the operational tests to ensure that variation in item characteristics will be limited. One such mechanism is the test sensitivity review (Hunter & Slaughter, 1980), which is intended to eliminate possible offensive items from the test. We are not suggesting that offensive items must be introduced into operational tests; we are suggesting that the screening of items should be accompanied by an empirical evaluation of its effects. Surely such screening has a value, indeed is essential in the modern political climate. But to apply it so rigidly that there is no opportunity to evaluate the reality of any anticipated consequences to test behavior may not be the best course, because the study of differential item performance is worth pursuing for several reasons. First, a continuing effort to produce better tests is generally regarded as desireable. For example, where test items handicap a particular group by introducing content that is essentially irrelevant to the characteristic being measured, that content should be discovered, modified, eliminated or counterbalanced with content that works in the opposite direction. Indeed, fairness demands that extraneous influences be eliminated or balanced. But if no

differential effects of consequence can be discovered and the intuitively obvious hypotheses fail, the documentation of the research that yielded the null results moves the burden of proof to the critics of the tests. Second, the test population can be conveniently partitioned using on a variety of demographic characteristics because there is considerable demographic variation in the examinee population. Fremer (1981), for example, has suggested studying differential item performances of rural and urban examinees--a partitioning for which he has some anecdotal data and one that is otherwise unexamined. Third, differential item performance studies have logistic value for developing scientific understanding. The items vary in many ways that are not described by subtest titles and can sample a wide variety of cognitive performances. The logistic mechanisms for doing differential item performance studies are convenient, as compared with other scientific manipulations, since they could be introduced as minor manipulations in the context of ongoing testing efforts; the results, if phrased in suitable terms, could lead to hypotheses, research and generalizations concerning other types of cognitive performances.

It should be mentioned that there has been some movement in the direction of hypothesis development. Levin (1970) has been quoted extensively to show how she departed from a set of hypotheses and arrived, through the examination of data, at a better understanding of the test performances of her NTE examinees. Donlon (1973) payed careful attention to the conceptual aspect of his work and was able to improve on the hypotheses formulated by Coffman (1961). Alderman and Holland (1981)

included the use of substantive experts in an attempt to develop conceptualizations. Francesco (1975) and Stricker (1981) both applied conceptualizations to all items and evaluated the results empirically. Scheunemann's (1981) ongoing study involves manipulating items in accordance with hypothesized characteristics to see if control over differential performance can be achieved. This latter study moves even further down the line of scientific development.

While we might not recommend all the details of Francesco's (1975) methodology, the general approach merits special emphasis. She obtained ratings of test items using judges who were not aware of differential item performance data. The items were rated on several characteristics; correlation and regression analyses used the ratings as independent variables, with item performance differences between groups taking the criterion side. Such studies have the following advantages: (1) they treat the items as having the characteristics in degree rather than as all or none phenomena, (2) all of the items contribute to determining the relationship between the characteristics and differential item performance, (3) the study creates defined characteristics that can be carried over from study to study rather than formulating them ad hoc, (4) the study anticipates future application by defining the judgments to be made, and (5) one set of data can, rather conveniently, be the subject of several rating studies that can lead to improved rating criteria.

Methodologies other than those on which the outlier studies are based are also needed, especially if one raises the possibility that the groups on which data have commonly been conditioned in the outlier are not most likely to be those whose use would lead to control of

differential performance. Perhaps it would be more effective to discover similar performing groups using only response data. Years ago Lazarsfeld confronted the problem of inferring a typology of examinees based on dichotomous data, which is indeed another way to approach the problem of differential item performance. As opposed to the recent outlier studies, which ask "Are there items that differ for certain groups in terms of their difficulties?", studies based on Lazarsfeld's latent class model (Green, 1951) would ask "Are there groups with different vectors of probabilities of correctly responding to the items?" Actually, if the Lazarsfeld question is revised to classify as group members those whose probability of correct responses are proportional, then the two questions become different sides of the same polyhedral solid. For unless more than one substantially sized group of the type Lazarsfeld sought exists, partitioning subjects to produce item difficulty plots shouldn't yield meaningful results regardless of the basis of the partitioning. If the existance of more than one group of appreciable size is indicated, the nature of the groups would be sought by examining item difficulties for the groups as well as demographic or other data. Such a study would no doubt have the not insignificant advantage of demonstrating that even though sex, ethnic and racial groups are related to the types of groups the relationship is not perfect, and hence we would be more reluctant to conclude that ethnic group membership "causes" performance differences.

Another side of the polyhedral solid is occupied by Donlon (1968) who has suggested that individuals' responses could be correlated with item difficulties to obtain a "personal biserial" correlation

coefficient. People who get easy items wrong and hard items right to an appreciable extent differ from the majority; if those with low personal biserials made, among themselves, similar responses they would constitute a group in the Lazarsfeld sense. Thus, as with the latent structure model, a product of the analysis could be the identification of groups with members who respond similarly. It should be mentioned that Harnish and Linn (1981) have discussed a whole family of indices that are related to the personal biserial; Tatsuoka and Linn (1981) have discussed other approaches including that of item response theory.

## References

Alderman, D. L., & Holland, P. W. Item performance across foreign language groups on a test of English as a foreign language. TOEFL Research Report, 1981, 9, Princeton, N.J.: Educational Testing Service.

Angoff, W. H. The use of difficulty and discrimination indices in the identification of biased test items. Berk (Ed.), Handbook of Methods for Detecting Item Bias. Baltimore, Md.: Johns Hopkins University Press, 1981.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W. H., Modu, C. C. Equating the scales of the Spanish-language Prueba de Aptitud Academica and the English-language Scholastic Aptitude Test of the College Entrance Examination Board. College Entrance Examination Board R&D Report RDR-72-73, No. 4. Princeton, N.J.: Educational Testing Service, Research Bulletin RB-73-4, January 1973.

Angoff, W. H., & Sharon, A. T. The evaluation of differences in test performance of two or more groups. Educational and Psychological Measurement, 1974, 34, 807-816.

Baker, F. B. A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 1981, 18, 59-62.

Birnbaum, A. Some latent trait models and their use in inferring on examinees ability, Part 5 in Lord and Novick.

Blew, E., & Ishizuka, T. College Board item bias study of the Scholastic Aptitude Test and the Test of Standard Written English--Form XSA4/E7. (ETS SR 78-62). Princeton, N.J.: Educational Testing Service, 1978.

Blew, E., & Stern, J. College Board item bias study of the Scholastic Aptitude Test and the Test of Standard Written English--Form XSA5/E8. (ETS SR 79-37). Princeton, N.J.: Educational Testing Service, 1979.

Breland, H. An investigation of cross-cultural stability in mental test items. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1974.

Cardall, C., & Coffman, W. E. A method for comparing the performance of different groups on the items in a test. College Entrance Examination Board Research and Development Report RDR-64-5, No. 9. Research Bulletin RB-64-61. Princeton, N.J.: Educational Testing Service, 1964.

Carlton, S. T., & Marco, G. L. Methods used by Educational Testing Service testing programs for detecting and eliminating item bias. Paper prepared for the 1980 Johns Hopkins University National Symposium on Educational Research entitled Test Item Bias Methodology: The State of the Art. Washington, D.C., 1980.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Coffman, W. E. Sex differences in responses to items in an aptitude test. 18th Yearbook of the National Council on Measurement in Education, 1961, 117-124.

Conrad, L., & Hallmark, M. M. Report on the item analysis of a GRE Aptitude Test by ethnic and sex subgroups. GRE Staff Paper, 1975.

Cook, L., & Stern, J. Item bias study of December 1974 SAT for black and white candidates. Princeton, N.J.: Educational Testing Service, unpublished memorandum, 1975.

Cowell, W., & Swineford, F. Comparisons of test analysis data for white male candidates with white female candidates. Report No. LSAC-72-2. In Law School Admission Council Report of LSAC Sponsored Research: Volume II, 1979-1974. Princeton, N.J.: Law School Admission Council, 1976.

Cowell, W. Special item analysis of the admission test for graduate study in business for candidates sponsored by the Consortium for Graduate Study in Business for Negroes. Princeton, N.J.: Educational Testing Service, ATGSB 690-10, 1968.

Donlon, T. Content factors in sex differences on test questions. Research memorandum RM-73-28. Princeton, N.J.: Educational Testing Service, 1973.

Donlon, T. F., & Fischer, F. E. An index of an individual's agreement with group determined item difficulties. Educational and Psychological Measurement, 1968, 28, 105-113.

Donlon, T. F., Ekstrom, R. B., & Lockheed, M. E. The consequences of sex bias in the content of major achievement test batteries. Measurement and Evaluation in Guidance, 1(4), 1979, 202-216.

Donlon, T. F., Hicks, M. M., & Wallmark, M. Sex differences in item responses on the Graduate Record Examination. Applied Psychological Measurement, 1980, 4, 9-20.

Echternacht, G. J. An examination of test bias and response characteristics for six candidate groups taking the ATGSB. Project Report PR-72-4. Princeton, N.J.: Educational Testing Service, 1972.

Flaugher, R. L., & Schrader, W. B. Eliminating differentially difficult items as an approach to test bias. Research Bulletin 78-4. Princeton, N.J.: Educational Testing Service, 1978.

Francesco, A. M. Item by sex interaction, an empirical investigation of the law school admfssion test. Report submitted to the Law School Admission Council, July 1975.

Fremer, J. Personal communication. Princeton, N.J., 1981.

Green, B. F. A general solution for the latent class model. Model of latent structure analysis. Psychometrika, 1951, 16, 151-166.

Harnish, D. L., & Fischer, F. E. Analysis of item response patterns: questionable list data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18(3), 133-146.

Hecht, L. W., & Swineford, F. Item analysis at Educational Testing Service. Princeton, N.J.: Educational Testing Service, 1981.

Humphry, B. J. A review od data based on the performance of a sample of black students from southern colleges on the NTE Common Examinations. Paper presented to the Consortium of Black Colleges, Houston, April, 1979.

Hunter, R., & Slaughter, C. ETS Test sensitivity review process. Princeton, N.J.: Educational Testing Service, 1980.

Levin, M. A review of the performance of black students in some humanities department tests. Princeton, N.J.: Educational Testing Service, unpublished memorandum, 1970.

Levin, M. Statistical properties of black literature items. Princeton, N.J.: Educational Testing Service, unpublished memorandum, 1970.

Linn, R. L., & Harnish, D. Intersections between item content and group membership on achievement test items. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1979.

Linn, R. L., & Levine, M. V., Hastings, C. N., & Wardrop, J. L. Item Bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 5(2), 159-173.

Lockheed, M. Sex bias in aptitude and achievement tests used in higher education. Princeton, N.J.: Educational Testing Service, 1981.

Lockheed, M. E. Sex bias in aptitude and achievement tests used in higher education, Chapter 5 in Perun, P. J., The Undergraduate Woman: Issues in Educational Equity. Lexington Books: D. H. Heath and Co., Lexington, Mass., 1982.

Lord, F. M. A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), Basic problems in cross-cultural Psychology. Amsterdam, the Netherlands: Swets and Zeitlinger B. V., 1977.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Addison Wesley, Reading, Mass., 1968.

Medley, D. M., & Quirk, T. J. Race and subject matter influences on performance on general education items of the National Teacher Examinations. Research Bulletin RB-74-43. Princeton, N.J.: Educational Testing Service, 1972.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Shepard, L., Camilli, G., Averill, M. Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Stratistics, 1981, 6,(4), 317-377.

Sinnott, L. T. Differences in item performance across groups. Research Report RR-80-19. Princeton, N.J.: Educational Testing Service, 1980.

Stern. J. College Board item bias study of the scholastic aptitude test and the test of standard written English. (SR-78-56). Princeton, N.J.: Educational Testing Service, 1978.

Stern, J. Item analysis and Delta plot study of SAT items for black and white candidates. Princeton, N.J.: Educational Testing Service, unpublished memorandum, 1974.

Stern, J. Item bias study of November 1974 SAT for black and white candidates. Princeton, N.J.: Educational Testing Service, unpublished memorandum, 1975.

Strassburg-Rosenberg, B., & Donlon, T. Content influences on sex differences in performance on aptitude tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Stricker, L. J. A new index of differential subgroup performance: Application to the GRE Aptitude Test. GRE Research Report No. 78-7. Princeton, N.J.: Educational Testing Service, 1981.

Tatsuoka, L., & Linn, R. L. Indices for detecting unusual item response patterns in personnel testing: Links between direct and item-response-theory approaches. Research Report 81-5. Computer-based Education Research Laboratory, University of Illinois, Urbana, Ill., 1981.

Wheeler, P., & Harris, A. Comparison of male and female performance on the ATP physics test. CB No. 4, College Entrance Examination Board, New York, 1981.

Wightman, L. E. Law School Admission Test: Comparisons of Canadian candidates with non-Canadian candidates (Report No. LSAC-76-5). In Law School Admission Council, Report of LSAC Sponsored Research: Volume III, 1975-1977. Princeton, N.J.: Law School Admission Council, 1977.